NATIONAL
CENTER for ANALYSIS of LONGITUDINAL DATA in EDUCATION RESEARCH

TRACKING EVERY STUDENT'S LEARNING EVERY YEAR

A program of research by the American Institutes for Research with Duke University, Northwestern University, Stanford University, University of Missouri-Columbia, University of Texas at Dallas, and University of Washington

AIR

Can UTeach? Assessing the Relative Effectiveness of STEM Teachers

Ben Backes
Dan Goldhaber
Whitney Cade
Kate Sullivan
Melissa Dodson

WORKING PAPER 173 • February 2018

Can UTeach? Assessing the Relative Effectiveness of STEM Teachers

Ben Backes
American Institutes for Research

Dan Goldhaber
University of Washington Bothell/American Institutes for Research

Whitney Cade
American Institutes for Research

Kate Sullivan
American Institutes for Research

Melissa Dodson
American Institutes for Research

Contents

Acknowledgements

The authors gratefully acknowledge financial support from the Institute of Education Sciences, U.S. Department of Education, through Grant R305A100116 to Southwest Educational Development Corporation, an Affiliate of American Institutes for Research. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. Neither AIR nor the authors of this paper have financial interests that could be affected by findings from the paper.

We thank the Texas Education Research Center for providing data. The conclusions of this research do not necessarily reflect the opinion or official position of the Texas Education Research Center, the Texas Education Agency, the Texas Higher Education Coordinating Board, the Texas Workforce Commission, or the State of Texas

CALDER • American Institutes for Research
1000 Thomas Jefferson Street N.W., Washington, D.C. 20007
202-403-5796 • www.caldercenter.org

Can UTeach? Assessing the Relative Effectiveness of STEM Teachers

Ben Backes, Dan Goldhaber, Whitney Cade, Kate Sullivan, Melissa Dodson

CALDER Working Paper No. 173
February 2018

Abstract

UTeach is a well-known, university-based program designed to increase the number of high-quality STEM teachers in the workforce. Despite substantial investment and rapid program diffusion, there is little evidence about the effectiveness of UTeach graduates. Using administrative data from the state of Texas, we measure the impact of having a UTeach teacher on student test scores in math and science in middle schools and high schools. We find that students taught by UTeach teachers perform significantly better on end-of-grade tests in math and end-of-course tests in math and science by 8% to 14% of a standard deviation on the test, depending on grade and subject.

Keywords: Teacher preparation; STEM; educator effectiveness

JEL codes: I20, I23

I. Introduction

A growing number of policymakers argue that for the U.S. to remain a worldwide economic and technological leader, it must do more to improve the quality of K–12 science, technology, engineering, and mathematics (STEM) education (e.g., Peterson et al., 2011). Given the growing body of evidence that educators are the most important determinant of student achievement outside of family and home influences (Borman & Dowling, 2008; Goldhaber, 2008; Hanushek et al., 2005; Heck, 2009; Ingersoll, 2001; Rice & Schwartz, 2008), it is no surprise that policymakers are focusing on teachers as a lever for improving STEM outcomes.[1] For instance, in fall 2009, President Obama asked his President's Council of Advisors on Science and Technology (PCAST) to draft a series of recommendations regarding the "most important actions that the administration could take to ensure that the United States is a leader in STEM education in the coming decades" (Holdren et al., 2010, p. vii). Among the council's findings was that math and science teachers are the "single most important factor in the K–12 education system…crucial to the strategy of preparing and inspiring students in STEM" (Holdren et al., 2010, p. 57).

The issue of STEM teachers is twofold. First, there are concerns about the quality of the existing STEM teacher workforce, particularly the prevalence of teachers without sufficient training in advanced subjects. For example, 61% of chemistry teachers and 67% of physics teachers do not hold a degree or certificate in those fields (Augustine, 2007). In addition, measures of overall science learning in the U.S. such as the National Assessment of Educational Progress (NAEP) find that science scores for high school students have shown no signs of improvement since 2009.[2] Second, there are longstanding issues of the quantity of STEM teachers and the difficulty of staffing STEM positions. Attracting STEM-trained

---

[1] Differences between assignment to effective versus ineffective teachers have been found to have profound impacts on students' test scores and later life achievement (Chetty, Friedman, & Rockoff, 2014a, 2014b; Clotfelter, Ladd, & Vigdor, 2007; Hanushek & Rivkin, 2010; Kane & Staiger, 2008).

[2] Garcia Mathewson, Tara. "NAEP: 4th, 8th grade science scores are up; 12th grade scores are flat." *Education Dive.* 27 October, 2016.

individuals to the teaching workforce is particularly difficult because of higher paying jobs outside of teaching (West, 2013), and between 2000 and 2012, 20-30% of schools reported difficulty filling STEM vacancies (Cowan et al., 2016).

UTeach is a relatively new program that is designed to address these quality and quantity issues by "transforming the way universities prepare teachers"[3] with an approach that the Obama Administration believes "has shown strong results".[4] President Obama's educational initiatives, such as Race to the Top, Change the Equation, and 100Kin10, place STEM teacher preparation at the center of national education reform efforts, and UTeach is featured in each of these programs. More recently it was recognized by the Obama Administration as a national model for increasing the number of teachers filling hard-to-staff positions in STEM (*Ed Week*, 2010).[5]

The UTeach program was created in 1997 by faculty at the University of Texas at Austin (UT Austin) in an effort to streamline the process of earning a degree in math or science alongside a teaching credential while graduating in a timely manner. Because of the perceived success of the program, it has spread rapidly. In 2014, the National Math and Science Initiative awarded a $22.5 million grant to continue the expansion of UTeach. Today it is available at 44 universities in 21 states, including state flagship universities such as UC Berkeley, the University of Florida, and West Virginia University, and is expected to produce more than 9,000 math and science teachers by 2020.

A selling point of UTeach is its approach to recruiting STEM majors to become teacher candidates while providing a pathway to have a STEM teaching credential in hand upon graduation. Students in UTeach take courses in their major along with classes for future teachers in a streamlined 4 year degree plan. Thus, UTeach has the potential to improve both the quantity and quality of the STEM

---

[3] *National Math and Science Initiative.* http://www.noia.org/wp-content/uploads/2013/03/40100.pdf
[4] https://www.whitehouse.gov/sites/default/files/docs/stem_teachers_release_3-18-13_doc.pdf
[5] Robelen, Erik W. 20 January, 2010. Obama Unveils Projects to Bolster STEM Teaching. *Education Week. http://www.edweek.org/ew/articles/2010/01/20/18stem_ep-2.h29.html*

teaching workforce by reducing barriers to entry to the teaching profession for STEM majors. As discussed in more detail below, the program touts three key elements that drive its success.[6] First, STEM majors are recruited as early as their freshman year; second, pedagogy courses are designed specifically for the program; and third, master and mentor teachers provide detailed guidance with early and intensive field experiences.

There are several reasons why UTeach teachers may be more effective at teaching STEM courses than the average teacher. First, by drawing from a pool consisting exclusively of math and science majors, the program potentially brings individuals with greater ability into the system than a typical teacher training program. On average, STEM majors who enter the teaching profession score about 100 SAT points higher than non-STEM majors (Goldhaber & Walch, 2013). Indeed, in the sample used in this study, the average replication site UTeach graduate scores 0.50 standard deviations higher on STEM certification exams than the average non-UTeach graduate. Second, subject-specific training may improve teacher performance in math and science at the secondary level, because some evidence suggests that greater math and science knowledge of teachers is associated with greater effectiveness – as measured by a teacher's ability to raise student test scores – at the high school level (Clotfelter, Ladd, & Vigdor, 2010; Goldhaber & Brewer, 1997; Goldhaber et al., 2016). Third, some UTeach-affiliated institutions such as UT Austin are more selective and thus may produce more effective teachers by this selection effect alone (Clotfelter et al., 2010).

Yet while selectivity of UTeach programs (UT Austin in particular) suggests that UTeach may be drawing more academically prepared individuals into the teacher workforce, there is reason to be cautious in thinking that this will necessarily result in significantly better student outcomes. For instance, the evidence on whether measures of college selectivity predicts teacher effectiveness is mixed (Harris

[6] http://www.senate.state.tx.us/75r/Senate/commit/c530/meetings/082304/downloads/Charge4_MRankin.pdf

& Sass, 2006).[7] Moreover, recent studies of traditional college and university-based teacher preparation programs (TPPs) suggests limited institutional level differences between TPPs (Goldhaber, Liddle, & Theobald, 2013; Koedel, Parsons, Podgursky, & Ehlert, 2015; von Hippel, Bellows, Osborne, Lincove, & Mills, 2016). In particular, Koedel et al. (2015) and von Hippel et al. (2016) emphasize that observed differences between TPPs are largely due to sampling variability rather than true differences between programs.[8]

In this paper, we use administrative data covering all math and science teachers and their students in public secondary schools in Texas to assess whether graduates of UTeach-affiliated programs in Texas are more effective than the average non-UTeach teacher as measured by student performance on standardized assessments. In doing so, we provide rare estimates of variation in STEM teacher quality in secondary schools, where subject-specific training may be most important (see also Clotfelter et al., 2010; Goldhaber et al., 2016; Jackson, 2014; Xu et al., 2011). We find that, relative to non-UTeach teachers in the state, UTeach-trained teachers are more effective as measured by their ability to raise student test scores in math and science. There are two important caveats: first, estimates for replication site UTeach graduates are not statistically significant for high school science or middle school math; and second, some results are sensitive to the decision of whether to include school fixed effects. However, as we describe, we also explore the sensitivity of the findings to the inclusion of school fixed effects and conclude that they are likely related to the sorting of UTeach teachers into schools that also tend to hire

[7] There is some evidence, for instance, that the positive findings for Teach For America teachers are partially explained by the selectivity of the Teach For America program (Xu et al., 2011); however, in their random assignment study, Clark et al. (2013) find that measures such as undergraduate selectivity and licensure scores do not explain any of the TFA effectiveness differential.

[8] Von Hippel et al. (2016) is especially relevant because, like this study, their sample consists of TPPs in Texas. However, while von Hippel et al. (2016) find little to no difference across TPPs in Texas, that does not guarantee that we will fail to find a UTeach effect because we have a much larger sample (five years of data compared to one), allowing for more precise estimates which could potentially mitigate the challenges imposed by sampling variability. In addition, one of our outcomes is performance on end-of-course exams, which are not included in the von Hippel et al. (2016) study and potentially allow for greater differentiation between TPPs due to the more advanced materials covered in EOC exams.

more effective teachers; i.e., the additional checks we perform tend to confirm the overall finding that UTeach teachers are more effective in general.

Based on our estimates, the difference between graduates from UTeach sites and non-UTeach teachers in the effectiveness with which they teach math courses is greater than the difference between novice teachers and teachers with 10+ years of experience in high school and is similar to the difference between novice teachers and teachers with 7 years of experience in middle school. While we find similar effects for Austin and UTeach replication site graduates in math, we find that in high school science, Austin graduates are substantially more effective than UTeach replication site graduates and other teachers in the state, which is partially but not fully explained by our measures of institutional selectivity, such as the SAT math scores of incoming students. Finally, while not the focus of the paper, we show descriptive evidence that the introduction of UTeach at partner universities has been associated with an increase in the number of STEM teachers produced.

II. UTeach Overview and Prior Research

UTeach introduced an approach that was not typically seen in higher education. UTeach undergraduate students can obtain a teaching certificate and graduate with a math or science degree in 4 years while taking teaching classes designed specifically for math and science teachers. The program streamlines content and pedagogy coursework to combine STEM degrees with secondary certification without adding time or cost to 4-year degrees. This feature is used as a recruitment strategy to attract STEM majors. According to internal data collected by UTeach, 55% of Austin UTeach graduates have graduated within 4 years, which is slightly higher than the university's overall average of 51-2%.[9] The

[9] UTeach average obtained through personal correspondence with Michael Marder, Co-Director of UTeach, 9 December 2016. Overall Austin average obtained through 2007 and 2009 entering classes from IPEDS: http://nces.ed.gov/collegenavigator/?q=Austin&s=TX&pg=2&id=228778

UTeach model has also been scaled up and replicated nationally in more than 40 universities with the support of both public and private funding.[10] A description of what UTeach sees as the key characteristics of its program follows.

Recruitment and Selection Strategies. Undergraduate STEM majors are recruited into the UTeach program as early as their freshman year with no selection criteria for entry. The program offers compact degree plans that allow STEM majors to complete their degrees and certification in 4 years. In addition, UTeach provides interested undergraduates with two one-credit-hour, field-based courses free of charge, allowing undergraduates to try teaching before committing to completing the teaching option.[11] Based on their experiences with these courses, undergraduates either choose to continue in the program or self-select out of the teaching option early in their college career.

*Preparation and Support for Preservice Teachers***.** In addition to the content courses required for their major, students who continue with the UTeach program complete a set of STEM-specific pedagogy courses that emphasize inquiry-based instruction, connections between the theory of the pedagogy and the practice of teaching, the interconnections between math and science, and the importance of diverse historical and methodological perspectives.

*Highly Structured Field Experiences***.** STEM majors enrolled in UTeach courses engage in approximately 40 hours of structured field experiences before student teaching, all of which are supervised by master teachers (non-tenured clinical faculty with prior teaching experience) and trained classroom mentor teachers. Before entering the student teaching semester, students are paired with local teachers who are trained to supervise and observe the UTeach student, offering multiple points for teacher candidates to reflect on their strengths and needs.

---

[10] These funders include the National Math and Science Initiative (NMSI), Exxon Mobil, Howard Hughes Medical Institute (HHMI), and state and federal resources.
[11] According to UTeach staff, between 2008 and 2012, 21% (lowest year) and 38% (highest year) of students who enrolled in this initial free field-based course at Austin eventually graduated from UTeach.

To guide program implementation at expansion sites, UTeach staff created the UTeach *Elements of Success,* a set of critical program components.[12] UTeach has also developed resources and support materials for all operational and instructional aspects of the UTeach model. Universities replicating UTeach receive direct and individualized support, including access to the *UTeach Operations Manual,* UTeach curriculum, student work samples, support materials, and support events, including course workshops and retreats, topical webcasts, the annual UTeach Conference, and UTeach Open House.

Research on UTeach has been conducted primarily by UTeach-affiliated faculty and their graduate students, with little or no third-party evaluation. Peer-reviewed studies of UTeach programs in the last 10 years have employed descriptive or correlational designs that relied on surveys, interviews, observations, reviews of student and teacher discourse, and reviews of transcripts, lesson plans, and other artifacts. One category of studies focuses on preservice and in-service teachers' knowledge, use, and perceptions of the efficacy of specific instructional approaches learned in the UTeach courses (Confrey, Makar, & Kazak, 2004; Dickinson & Summers, 2010; Marshall & Young, 2006). Another set of studies explores preservice teachers' development and use of mathematical and statistical discourse (Ares, Stroup, & Schademan, 2009; Makar & Confrey, 2005). Finally, studies by Stroup, Hills, and Carmona (2011) and Marder and Walkington (2014) focus on exploring statistical approaches and methods for analyzing administrative and qualitative teacher and student data. The latter finds, for example, that the classroom observation protocol developed for UTeach is fairly weakly correlated with value-added scores.

[12] UTeach *Elements of Success:* https://institute.uteach.utexas.edu/sites/institute.uteach.utexas.edu/files/uteach-elements-of-success-2011.pdf

III. Data

We use detailed student-level administrative data that link students in Texas to their teachers for five school years (2011–12 through 2015–16).[13] Texas has the second largest public K–12 enrollment in the United States, and large minority and disadvantaged student populations: about 52% of its students are Hispanic, 13% Black, and 30% White, and about 60% of students identified as economically disadvantaged.[14]

The student-level longitudinal data we use in the analysis contain math and science scores as primary outcome variables.[15] For math, these include both end-of-grade (EOG) and end-of-course (EOC) exams, with the bulk of EOC scores coming from Algebra I (77.5%) and the remainder Geometry (16.9%) and Algebra II (5.5%).[16] The share of Geometry and Algebra II students is relatively small because those tests were administered for only two of the five years some years that our sample covers. For science, we include EOC scores as outcome measures (81.8% Biology, 15.4% Chemistry, and 2.8% Physics), with Grade 8 EOG science used as a control for students in ninth grade.[17] Although estimating models with science test scores as an outcome variable in value-added models has not been as thoroughly vetted as math and reading, we perform a number of robustness checks, such as controlling for eighth-grade science scores for upper-grade students and estimating UTeach effects on a sample of schools that do not seem to group students of similar ability by classroom. Our results for EOC science are similar when

[13] We also use test score data from 2010–11 as prior test scores for regressions using the 2011–12 data for the outcome measure.
[14] tea.texas.gov/acctres/Enroll_2013-14.pdf
[15] In the 2010–11 school year, all students between grades 3 and 11 took the TAKS in math. However, with the introduction of EOC exams in 2011–12, the STAAR test began to be phased in with the 2011-12 school year. Since then, STAAR has been administered in Grades 3–8 with additional EOC exams in Algebra I and Biology. In 2012 and 2013 only, there were also EOC exams for Algebra II, Geometry, Chemistry, and Physics. For students taking an EOC exam, we consider their most recent test score to be their lagged test score. A Grade 8 science EOG test has been administered since 2012. See http://tea.texas.gov/student.assessment/staar/ for more information.
[16] We also use students' prior reading test scores as a control variable.
[17] We do not use EOG science scores as an outcome variable because this test is not administered to sixth- or seventh-graders. EOG (i.e., 8th grade) science scores are only used as control variables in regressions where EOC science scores are the outcome.

performing these checks. Finally, students' scale scores are standardized to have mean 0 and standard deviation 1 at the subject-grade-year-test level within the state.

In addition to standardized test scores, we observe a variety of student characteristics: race/ethnicity, gender, free- or reduced-price lunch (FRL) eligibility, gifted status, limited English proficiency (LEP) status, and disability status, which are used as covariates in our analyses. In addition, all students are linked to teachers based on course enrollment.[18]

Teacher personnel files contain information on teachers' experience, undergraduate and graduate institutions, demographics, and other supplemental background variables. These are likewise used as covariates for some of the models in the analysis that follows. UTeach teachers are identified by combining where each teacher earned his or her degree, graduation year, and subject of teaching certificate. According to UTeach, it is possible to obtain undergraduate training to become a STEM-certified teacher from UTeach universities only by going through the UTeach program.

We begin by describing placement patterns of UTeach graduates by UTeach site, year, and subject; subjects include EOG math (Grades 6–8), EOC math (Algebra I, Algebra II, and Geometry), and EOC science (Biology, Chemistry, and Physics). A teacher is counted in a given sample if he or she teaches a student in that sample; thus, it is possible for teachers to appear in multiple samples. For example, if a teacher taught both eighth-grade math and Algebra I, then he or she would appear in both the EOG math and EOC math samples. We display placement patterns in this manner because we obtain estimates separately for these three samples.

Counts of UTeach teachers by campus and year are shown in Table 1. Two patterns are readily apparent. The first is that the number of teachers from UT Austin is relatively steady over time, whereas the number of teachers from the replication sites grows substantially. Austin's program dates back to

[18] Teachers of record in students' core math and science courses are linked to them for the analysis. Student observations linked to multiple teachers (e.g., due to coteaching, student mobility) are weighted in proportion to the amount of time spent with each teacher, based on available enrollment data (Hock & Isenberg, 2012).

1997, well before the coverage of our data, suggesting the number of UTeach teachers graduating from Austin has more or less stabilized, whereas many of the other sites began their UTeach replication relatively recently and are ramping up graduate numbers. For example, although the number of UTeach graduates in EOC math classrooms from Austin was virtually the same in 2012 and 2016 (115 teachers vs. 120), the number increased from 12 to 51 at Houston and from 12 to 80 at North Texas during the same time frame. Second, UTeach graduates are concentrated in EOC subjects instead of the EOG grades. In 2016, for example, only 18 Austin UTeach graduates taught students who took EOG tests, while 120 taught math and 82 science EOC subjects, respectively.

While this paper is primarily interested in the question of whether UTeach programs produce more effective teachers than other TPPs, given the above discussion about the difficulty of staffing STEM positions, it is also important to consider whether the introduction of UTeach is also associated with an increase in the number of STEM teachers produced. Table 2 displays the number of teachers who appear in our analysis sample by campus and graduation calendar year. At the two replication sites whose first UTeach graduates finished in 2010, Houston and UNT, we see substantial increases in the number of STEM teachers. At Houston, for example, the three years prior to UTeach saw 10, 5, and 5 teachers enter the workforce (average of 7 per year), while in the three years after, there was an increase to 10, 11, and 27 teachers (16 per year). At UNT, the increase was even larger, from 19, 19, and 13 teachers (17 per year) to 30, 29, and 38 teachers (32 per year). Thus, while not conclusive, this descriptive evidence is consistent with UTeach fulfilling its goal of recruiting more students to become STEM teachers.

Table 3 presents descriptive statistics of the students taught by UTeach and non-UTeach teachers included in the study. As with the counts discussed above, the samples used here are, of necessity, limited to grades and subjects in which standardized tests are administered to students. We keep the same groupings of teachers by EOG math, EOC math, and EOC science, while also splitting the sample into Austin UTeach, replication (non-Austin) UTeach, and non-UTeach teachers. We choose this grouping for two reasons. First, to the extent that Austin UTeach produces teachers that are more effective on average, an important question is whether that success can be re-produced at other campuses. Second, because the replication sites do not have sufficient observations to obtain campus-by-campus estimates with any degree of precision, we group them together to measure a collective replication site effect.

In the EOC courses where Austin graduates are concentrated, for Austin graduates relative to non-UTeach graduates, the students of Austin graduates are similar in EOC math but more advantaged in EOC science. In EOC science, Austin graduates are less likely to teach FRL-eligible students and more likely to teach gifted students and students whose prior achievement was substantially higher.[19] On the other hand, UTeach graduates from the replication sites are more likely to teach black students, LEP students, and students with lower prior achievement, compared to graduates from Austin.[20]

Teacher characteristics of UTeach and non-UTeach teachers are shown in Table 4. The typical UTeach teacher—whether from Austin or not—has fewer years of experience than the typical non-UTeach teacher. This is especially pronounced for the non-Austin group because, as discussed previously, these programs are relatively new: in all subjects, the percentage of teachers from replication sites in their first through third year of teaching is about 80%. In addition, the average selectivity of the undergraduate institution attended is higher for both UTeach samples (Austin and replication) —dramatically so for Austin. UTeach graduates are also more likely to be present in the certification database as STEM-certified; in fact, all UTeach teachers in our sample have a STEM

[19] Prior math scores do not average 0 across all students in the EOC math sample due to selection into advanced math courses (test scores are standardized among all students in the state). For example, if students progress to Geometry only if they have sufficiently high Algebra I scores, then the prior test scores for students taking Geometry will be higher than the mean of zero because of this selection mechanism. UTeach estimates with EOC scores as the outcome are similar when including 8th grade EOG math and reading scores as additional controls.
[20] The patterns for school-level measures of student demographics and ability for schools with UTeach present vs. all schools are similar to the student-level patterns in Table 3.

certification because of the way we construct the UTeach variable, in which a teacher has to graduate from a UTeach campus with a STEM certification. In contrast, more than one third of EOG math teachers are either not in the certification database or do not have a STEM-specific certification. Finally, math UTeach teachers are substantially more likely to be Hispanic than the average teacher in Texas.

IV. Methods

Our baseline analysis measures the difference in relative effectiveness between UTeach-trained teachers and comparison teachers who teach math or science to secondary students. Our approach follows similar studies of individual TPPs such as Teach For America and the New York City Teaching Fellows Program (Boyd, Lankford, Loeb, & Wyckoff, 2006; Hansen, Backes, Brady, & Xu, 2015; Kane, Rockoff, & Staiger, 2008). We estimate the following equations:

$$y_{ist} = \beta_0 + \beta_1 y_{ist-1} + \beta_2 X_i + \beta_3 UTeach_i + \beta_4 T_j + \varepsilon_{ist} \text{ , (1)}$$

where y_{ist} indicates the score on an EOG or EOC math or science exam (with separate regressions for each) for student *i* in school *s* in year *t*, y_{ist-1} a vector of cubic functions of prior year test scores in math and reading (and science, when science performance is the outcome measure), $UTeach_i$ is an indicator for whether student *i* was taught by a UTeach graduate in the tested subject, X_i contains a vector of student *i*'s characteristics, including race, gender, eligibility for FRL, special education status, and gifted status, and T_j a vector of controls for teacher characteristics, which in most models consists solely of experience. Students with missing prior year scores are assigned a value of 0 for prior score with an additional control for missing prior year scores.[21] For students who took multiple tests in the prior year (e.g., Algebra I and EOG eighth-grade math), we use EOG scores as the measure of prior year

[21] EOC results for students in higher grades (10 and above) are similar when controlling for eighth-grade EOG scores in addition to prior year EOC scores.

achievement in regression models. In models where EOC tests are pooled together as an outcome variable (e.g., EOC math score as the outcome variable), we interact subject with all control variables to allow the association between these variables and the outcome to vary by test type. In addition, ε_{ist} represents a randomly distributed error term. In all analyses, standard errors are clustered at the teacher level.

The coefficient of interest, β_3, represents the average differential effectiveness of UTeach graduates relative to other teachers in the state. Both experimental work and nonexperimental tests suggest that controlling for prior test scores as in Equation (1) is sufficient for estimating teacher effects with little bias (Bacher-Hicks, Kane, & Staiger, 2014; Chetty et al., 2014a; Kane, McCaffrey, Miller, & Staiger, 2013; Kane & Staiger, 2008), with the caveat that these studies do not examine high school teachers. Obtaining unbiased estimates of the effect of certain teacher characteristics on student achievement at the high school level is likely to be more challenging given the greater prevalence of specialty high schools and ability tracking (Jackson, 2014). We attempt to account for the potential that students with unobserved attributes correlated with test achievement are tracked into schools or classes (Jackson, 2014) by estimating additional models that include school or track effects. In these models, the effects of UTeach teachers are identified based on comparisons within the same school or track, where a track is defined to be all students within the same school who take the same set of courses in the same year.

To account for school effects, we estimate models where we add school fixed effects to Equation (1) and thus compare UTeach graduates to other teachers within the same school rather than to all non-UTeach teachers in the state. In these specifications UTeach-trained teachers are compared to teachers in similar school settings. However, as discussed in further detail below, this approach is potentially problematic because school fixed effects could absorb true differences in teacher effectiveness across schools. For example, if it were the case that UTeach teachers and the comparison teachers in UTeach schools were all truly more effective on average, adding school fixed effects would difference out some of the true effectiveness of UTeach teachers by comparing them to more effective teachers. In order to investigate the possible extent of differential teacher sorting into UTeach schools, we estimate the following model:

$$CERT\_SCORE_{ist} = \beta_0 + \beta_1 y_{ist-1} + \beta_2 X_i + \beta_3 SCH\_UTeach_{is} + \beta_4 T_j + \varepsilon_{ist} \text{ , (2)}$$

where $CERT\_SCORE_{ist}$ is the certification score of student *i*'s teacher and SCH_UTeach_{is} is an indicator for whether student *i* is in a school that ever employed a UTeach teacher. As with Equation (1), we estimate Equation (2) subject-by-subject, so SCH_UTeach_{is} denotes ever hiring UTeach in the subject under consideration in order to obtain estimates of the potential degree of teacher sorting by subject. Because it is the comparison teachers in UTeach schools that are of interest, we omit UTeach teachers from all models that estimate Equation (2). In Equation (2), the coefficient β_3 measures the extent to which a student being in a UTeach school is differentially predictive of being exposed to teachers with higher certification scores, controlling for student background and prior achievement in the same manner as our basic model in Equation (1). In addition, we also estimate the following model for students and teachers who are not in a UTeach school in a given year:

$$y_{ist} = \beta_0 + \beta_1 y_{ist-1} + \beta_2 X_i + \beta_3 Ever\_SCH\_UTeach_{ist} + \beta_4 T_j + \varepsilon_{ist} \text{ , (3)}$$

where $Ever\_SCH\_UTeach_{ist}$ is an indicator for whether student *i* is being taught by a teacher who would ever teach at a UTeach school.[22] In Equation (3), the coefficient β_3 measures whether being taught by a teacher who would ever teach in a UTeach school is associated with differential student achievement, conditional on student demographics and prior achievement.

[22] We control for teacher experience in Equation (3) to account for the possibility of teachers being in different stages of their careers when they teach in UTeach schools relative to non-UTeach schools.

Equation (1) yields an estimate of the average difference in achievement between students who were taught by UTeach graduates and those who were not. To investigate heterogeneity across programs, we decompose the UTeach coefficient into separate coefficients for different campuses (UT Austin, University of Houston, University of North Texas, UT Dallas, UT Arlington, and UT Tyler) to assess whether different UTeach sites produce teachers of varying effectiveness:

$$y_{ist} = \beta_0 + \beta_1 y_{ist-1} + \beta_2 X_i + \sum_j \alpha^j \, Campus_i^j + \beta_4 T_j + \varepsilon_{ist} \text{ , (4)}$$

In Equation (4) above, α^j represents the coefficient estimate for each separate UTeach campus *j* and measuring the average difference between campus *j* and the average non-UTeach teacher. Variation in the α^j coefficients would indicate the extent to which graduates of different UTeach sites are differentially effective. It could be the case, for example, that graduates trained at the founding site, UT Austin, are more effective than graduates from the replication sites as a result of higher implementation fidelity (as the founding site) or because UT Austin is the most selective of the UTeach campuses. We test for this explicitly by grouping UTeach schools into Austin and replication (non-Austin) campuses.[23]

An important question is the extent to which UTeach effects are driven by the UTeach program itself rather than general institution or selectivity effects. For example, as of 2014, 41% of UTeach graduates nationwide were trained at UT Austin, ranked as a "highly competitive" university in Barron's Profiles of American Colleges.[24] In their study of Teach for America (TFA), a selective teacher training program, Xu et al. (2011) found that a substantial portion of the greater effectiveness of TFA instructors relative to other teachers can be explained by TFA's selection of candidates with better observable characteristics, such as graduating from a more selective university and having higher licensure test (Praxis) scores. Thus, one may expect UTeach teachers from UT Austin to be the most effective, given

[23] In practice, there are not enough observations for each campus to obtain informative estimates for individual campuses based on Equation (2), so we group UTeach campuses into Austin and replication (non-Austin) sites for most of the analyses described below.

[24] UTeach program data: "UTeach and UTeach Expansion" from uteach-institute.org.

that UT Austin is more selective than the other UTeach campuses. We thus perform a series of tests to investigate the question of program versus institution effects. First, we investigate whether selectivity can explain the UTeach effect by exploring the sensitivity of results to the addition of selectivity measures at the campus (SAT scores of incoming freshman students) and teacher (licensure scores) levels. Second, for replication sites, we compare the relative performance of UTeach graduates to those who graduated from those same institutions prior to the introduction of UTeach to see if the introduction of UTeach was associated with an increase in the performance of graduates from a given replication site.

V. Results

Before displaying our estimates of UTeach effectiveness, we first note two ancillary findings that place our UTeach estimates in context. First, we estimate the dispersion of teacher effects by subject – sometimes referred to as the teacher "effect size" – by estimating models with teacher fixed effects and shrinking these estimates using an Empirical Bayes procedure. The standard deviations for these teacher fixed effects for each subject are as follows: EOG math 0.22, EOC math 0.44, EOC science 0.29, EOG reading 0.14, and EOC reading 0.27, with the results for the math and science subjects here being consistent with Goldhaber et al.'s (2016) estimates from the state of Washington. Second, using Lipsey et al.'s (2012) estimates of annual learning by subject and grade, taking the most conservative (i.e., largest standard deviation) estimates for translating test score gains to months of learning, the average student gains 0.32 standard deviations per year in middle school math, 0.25 standard deviations per year in high school math, and 0.22 standard deviations per year in high school science.[25]

[25] In all months of learning calculations, we use a 9 month school year.

A. Baseline Findings

(i) UTeach Estimates

We begin by displaying the results of our basic estimating equation for math subjects in Table 5. Each math test—EOG math, Algebra I, Algebra II, and Geometry—is used as an outcome variable in four specifications, with each column representing the results from a different specification.[26] The first three columns show results with no fixed effects but different choices for controls, and the last column contains school fixed effects. Results are mostly consistent across the different tests: without fixed effects, the typical student in a UTeach classroom scores 0.05-0.13 standard deviations higher, depending on the subject and model.

When adding fixed effects, there is a reduction in effect size from 0.05-0.13 to 0.01-0.14 standard deviations. Without Algebra II, which has the smallest sample size of UTeach teachers, the fixed effects range is 0.01-0.08. As we discuss in greater detail below, is possible that the shrinkage of the UTeach effect size when adding school fixed effects (whether at the school level or school-track level) is due to teachers of similar effectiveness sorting into the same schools (e.g., Goldhaber et al., 2013). This is the opposite of the result sometimes observed in TFA evaluations, in which the TFA effect increases with school fixed effects because the comparison teachers in the disadvantaged schools where TFA corps members are placed tend to be below average (e.g., Hansen et al., 2015).

We display the same models for the science subjects (Biology, Chemistry, and Physics) in Table 6 with similar results, albeit consistently larger in the models without school fixed effects. Although the ordinary least squares (OLS) results for Physics in Table 6 are very large, this constitutes a small share of the sample, and this reflected by the large standard errors.

[26] Appendix Table 1 shows selected coefficient estimates from column 2 of the three most common subjects: EOG math, Algebra I, and Biology. For the sake of space we do not report the coefficients in the main tables of the paper.

The effect sizes shown in Tables 5 and 6 are large relative to the returns to teacher experience and to how much a typical student learns in a year. In EOG math, column 2's 0.10 standard deviations is similar to the difference between a teacher with 0 years of experience and one with 7 years of experience (see Appendix Table 1). In a 9 month school year, 0.10 standard deviations translates to 2.8 months of additional learning. For Algebra I and Biology, the UTeach effect is larger than the difference between teachers with more than 10 years of experience and teachers with 0 years of experience and is equivalent to 4.7 months of learning in math and 3.7 months in science. Although our estimates of returns to experience may appear small (0.06 for teachers with more than 10 years of experience in Algebra I and 0.05 in Biology), they are not substantially different than other estimates of the returns to experience using teachers in high school subjects (e.g., Clotfelter et al., 2010, Xu et al., 2011).

(ii) Investigating the Differences Between the OLS and School Fixed Effect Models

The results in Tables 5–6 indicate that the decision whether to include school fixed effects leads to meaningful differences in the magnitudes and statistical significance of the estimates in some subjects. In models without school fixed effects—thus comparing UTeach graduates to all other STEM teachers of a given tested subject in the state—UTeach teachers perform better than the average non-UTeach teacher in all subjects. However, when comparing UTeach graduates to other STEM teachers in the same subjects in the same schools, the estimates for many subjects are attenuated and lose statistical significance.

As discussed in Goldhaber et al. (2013), it is not obvious which set of results should be privileged. On one hand, the fixed effects model is theoretically attractive because it removes potential time-invariant biasing factors such as principal quality, curriculum, or teacher collegiality. On the other hand, if UTeach graduates sort into schools where both UTeach and comparison teachers are truly more effective, then the addition of school fixed effects obscures the true effectiveness of UTeach teachers by restricting the comparison group to teachers who are more effective than average.

Our basic results in Tables 5 and 6 show that even with school fixed effects, UTeach is estimated to be more effective in all subjects except for Algebra I and Biology (although not always statistically significant in all subjects). In addition, as we will show below, EOC science results are positive and statistically significant for Austin graduates and not statistically significant for replication site graduates whether or not fixed effects are included. Thus, understanding why our estimates are sensitive to the inclusion of school fixed effects is most important in Algebra I, especially because this is the subject with by far the greatest number of UTeach teachers.

We conduct two additional analyses to investigate whether the sensitivity of the estimates may be related to the sorting of teachers into schools employing UTeach teachers. First, we measure the certification scores of the comparison teachers in UTeach schools; and second, we examine the performance of students assigned to the comparison teachers in the years in which comparison teachers were not teaching in UTeach schools. As we describe below, we believe these tend to support idea that the UTeach fixed effects estimates are attenuated due to comparison teachers being more effective, especially in Algebra I.

The first test measures the certification scores of students' teachers, which are positively correlated with student achievement in math and science both in our data and in other settings (see, e.g., Clotfelter et al., 2010, and Goldhaber et al., 2016). In Panel 1, we provide measures showing the degree to which a student being a UTeach school in a given subject is associated with being taught by teachers with higher or lower certification scores, conditional on student background and prior achievement (because it is the comparison teachers in UTeach schools that are of interest, these regressions omit UTeach teachers). Panel 1 shows that UTeach schools are more likely to contain students taught by comparison teachers with substantially higher certification scores. In the high school subjects, these estimates range from 0.07 standard deviations in Algebra II to 0.16 standard deviations in Geometry. In Algebra I and Biology, the two subjects with the largest UTeach sample sizes that are the most sensitive to the inclusion of school fixed effects, the UTeach school coefficients are statistically significant and about 0.10 standard deviations higher, suggesting that the attenuation of UTeach coefficients with the addition of school fixed effects being driven by stronger within-school comparison teachers is plausible.

The second test investigates the performance of UTeach comparison teachers in the years in which they taught outside of UTeach schools. It is important to look at these teachers outside of UTeach schools because otherwise we have the same fundamental problem of measuring teacher versus school effects that we have with UTeach teachers, where we cannot disentangle school effects from teacher effects. If it were the case that the comparison teachers in UTeach schools are truly more effective, we would expect that, among non-UTeach teachers who taught in both UTeach and non-UTeach schools, they would also be more effective in non-UTeach schools. Because the coefficient of interest is identified based on teachers who taught in multiple schools (i.e., both UTeach and non-UTeach schools), we display results for the three subjects where we observe more than two years of data: EOG math, Algebra I, and Biology. Panel 2 of Appendix Table 2 shows the results from estimating Equation (3), a regression of student achievement on an indicator for whether a teacher ever taught at a UTeach school, *in a sample restricted to non-UTeach schools*, controlling for students' demographic information and prior test scores. The results indicate that being taught by a teacher who would at one point be present in a UTeach school predicts in increase of student learning in Algebra I. While there is no difference in EOG math or Biology, our results are not sensitive to the inclusion of school fixed effects in EOG math.

Finally, given the findings from the teacher value added literature, it would be surprising if the UTeach effects we observe were driven entirely by the schools in which they teach rather than the underlying ability of teachers. The value added literature is relevant because the UTeach estimates in the paper are essentially what would be obtained if we obtained individual estimates for the value added of every UTeach teacher by estimating teacher fixed effects and averaging them together, weighted by the number of students they taught. A number of different papers (e.g., Chetty et al., 2014; Xu et al., 2012) show that the achievement of the students taught by teachers who are entering a new school can be predicted by the achievement of the students taught by those teachers in their prior schools, even when switching between schools with substantially different poverty or achievement levels. An important caveat here, however, is that these results have largely been obtained from students in grades 4-8, while in our setting it is the end of course results that are most sensitive to the inclusion of school fixed effects.

In the remainder of this paper, we display estimates both with and without fixed effects in the interest of transparency but largely focus the discussion on the models without fixed effects due to the potential problems with fixed effects models noted above.

B. Accounting for Teacher Characteristics

The first three columns of Tables 5 and 6 control for different teacher characteristics. Column 1 includes no teacher characteristics. Column 2 adds teacher experience. In all seven tests used as outcomes, including teacher experience increases the magnitude of the UTeach estimate. This is not surprising because UTeach teachers have less experience on average than other teachers, so accounting for this experience differential makes UTeach performance look relatively stronger.

One potential explanation for the effectiveness of UTeach teachers relative to comparison teachers is that all UTeach teachers are STEM-certified, while about one third of EOG teachers and 5 to 10 percent of EOC teachers are either certified in a non-STEM-specific field or absent from the state’s certification database. Although the literature presents mixed evidence about the relationship between certification and student achievement (e.g., Rockoff, Jacob, Kane, & Staiger, 2011), we nevertheless explore how the addition of control for STEM certification would affect our UTeach estimates. Results are shown in column 3 of Tables 5 and 6. Interestingly, the STEM certification coefficients for Algebra I and Geometry in column 3 (0.36 and 0.25, respectively; not shown but available from authors) is larger than the corresponding result for Algebra and Geometry teachers in North Carolina (0.13) from Clotfelter et al. (2010), but the association between Biology score and STEM certification in our study (0.04) is similar to Clotfelter et al. (2010)'s result (0.03).[27] Returning to UTeach estimates, relative to columns 1 and 2 of Tables 5 and 6, the coefficients for UTeach in all math aside from Algebra II are modestly attenuated when including STEM certification as a control. However, even with these controls, the coefficients remain statistically significant and positive. For EOC science, because only 5% of non-UTeach teachers are missing a STEM credential in our data, adding controls for STEM certification does not affect the UTeach coefficient. Thus, providing a guided route to certification may explain a small portion of the effectiveness of UTeach in math classrooms but UTeach teachers remain more effective than the average teacher in Texas even after accounting for STEM certification.

C. Results for Austin vs. UTeach Replication Sites

Next, we present separate results for the two types of UTeach campuses: the original site at Austin and all other replication sites at universities in Texas, as shown in Table 7. Because the results from Tables 5 and 6 are generally consistent across EOC tests within the same subject for the subjects where most students are concentrated, for the remainder of the results section we group results into

[27] We do not show the coefficients in the tables for sake of brevity. They are as follows: EOG M 0.04, Algebra I 0.36, Geometry 0.25, Algebra II 0.13, Biology 0.04, Chemistry 0.15, Physics 0.04. Note that some of these coefficients have changed from an earlier version of the paper where certification status was missing for a greater share of teachers due to incomplete data, resulting in artificially low STEM certification rates.

the three categories used in the summary statistics tables: EOG math, EOC math, and EOC science.[28] For EOC math, observations largely come from Algebra I (75% of observations), whereas for EOC science, Biology (80%) makes up the bulk of observations.

The results in Table 7 indicate that achievement gains associated with UTeach classrooms are similar for Austin and replication site graduates in math (panels 1 and 2) but that Austin graduates are substantially more effective in science (panel 3). In column 2, The EOC math achievement boost is estimated to be 11% of a standard deviation for Austin graduates (4.0 months of additional learning) and 14% for replication site graduates (5.0 months). The difference between Austin and replication site graduates in EOG math and EOC math is not statistically significant in any model that controls for teacher experience. In science, however, the estimated achievement gain for students of Austin UTeach graduates is 13% of a standard deviation (5.3 months) without school fixed effects or controls for STEM certification (column 2) and 4% with school fixed effects (column 4). For the non-Austin replication sites, the corresponding gains are 4% in EOC science (1.6 months) without fixed effects, and below zero with fixed effects, with neither being statistically significant. When disaggregating non-Austin sites by campus, shown in Appendix Table 3, standard errors become too large to be confident in the findings for most sites, although many point estimates for UT Dallas and Houston are very large and positive.

D. Robustness Tests and Other Checks

In Table 8, we display the results from a series of robustness checks, with rows 1 and 2 showing our main results from Table 7 for comparison. Row 3 replaces school fixed effects with school-track fixed

[28] In these specifications, we interact coefficients by test type to allow their association with the outcome to vary by test type. For example, in Table 5, in results not shown, prior math scores are differentially predictive depending on the outcome test under examination.

effects (Jackson, 2014).[29] Results indicate that the choice between school fixed effects (row 2) and school-track fixed effects (row 3) makes almost no difference for the high school subjects where tracking is the strongest concern, although the replication site EOG math results are substantially attenuated.

In rows 4 and 5, we add school characteristics (row 4) and school plus class characteristics (row 5) to the basic OLS model.[30] Relative to row 1, this results in some attenuation; however, the results are qualitatively similar in that UTeach teachers are still seen to be more effective. As discussed in Goldhaber et al. (2018), it is difficult to know whether the attenuation when classroom controls are added is due removing true differences across classrooms by over-controlling for the influence of peer effects or if the attenuation reflects less biased estimates.

In rows 5 and 6, we restrict our sample to schools that do not appear to track by ability grouping (about 80% of schools),[31] UTeach coefficient estimates are smaller for Austin graduates teaching EOC math (0.06 without school fixed effects) but similar for Austin graduates teaching EOC science and for replication site graduates in all subjects.

To this point, our sample has included all teachers regardless of the amount of time elapsed since they completed their training. If it were the case that TPP effects are strongest for teachers who recently graduated and entered the workforce and that they dissipate as time elapses, as found in Goldhaber et al. (2013), one would expect the UTeach effects estimated above to be an attenuated version of what might be found on a sample of novice teachers. In rows 8 and 9, we run our basic models on a sample of teachers in their first, second, or third year of teaching. Three of the four EOC coefficients are larger when the sample is restricted to novice teachers, but the coefficients are smaller

[29] Depending on the subject, 10-20% of school-tracks with UTeach have no variation in UTeach (i.e., the only students in that track are exclusively taught by UTeach in that subject), meaning that these school-tracks no longer contribute to our UTeach estimates.
[30] Specifically, these include cubic polynomials of school and/or class averages of percent black, percent Hispanic, and prior math and reading scores.
[31] Specifically, we retain a sample of schools where the average deviation in prior math test scores at the classroom level from the overall school average is less than 0.05 standard deviations.

for EOG math. For example, for EOC math, the coefficient increases from 0.11 to 0.13 for Austin and 0.14 to 0.15 for replication sites after the sample restriction. This pattern is consistent with TPP effects being most pronounced for teachers who have recently completed their training for teachers in advanced subjects.

Rows 10 and 11 perform a robustness check in which we drop the first cohort from each UTeach site. If it were the case that some graduates of a university's pre-UTeach program were still finishing their program in the early stages of UTeach, dropping the first cohort would avoid misclassifying these teachers as UTeach. Results are broadly similar.

To sum up Table 8, EOG math results are always positive but rarely significant for Austin and replication sites, EOC math results are positive and significant for nearly all models that do not include fixed effects for Austin and replication sites, and EOC science results are always positive and significant for Austin but have mixed sign and no significance for replication sites.

E. Exploring Heterogeneity Across UTeach Sites

In this section, we investigate potential explanations for why the observed achievement gains associated with being in a UTeach classroom are stronger for the Austin graduates in science than for graduates from the replication sites. We run through three potential explanations that could plausibly explain the Austin-replication site differential. First, we investigate institution-level selectivity measures (SAT scores of incoming students); second, we test individual-level ability measures (teacher certification scores); third, we investigate whether the Austin effect is driven by its program having been in place longer by looking for evidence of the replication sites improving over time. Before proceeding, we emphasize that we cannot cleanly distinguish selection effects from training effects. Thus, when we measure an "Austin effect," we are measuring the result of the combined process of the sorting of students into Austin, the sorting of Austin students into training, and the training itself.

The first two potential explanations pertain to selection. We experiment with two ways of accounting for selection and then discuss them together. First we control for the selectivity of the undergraduate institution attended, as measured by the 75th percentile of math and reading SAT scores of incoming freshmen. Specifically, we add dummy variables for each selectivity quintile for both math and reading SAT scores to allow for potential nonlinearities between selectivity and effectiveness, with an additional dummy indicating missing scores (either because the undergraduate institution does not require SAT scores for admission or because we cannot identify in the state database which institution a teacher attended).[32] Although SAT scores are only one dimension by which students are selected into college, they are nevertheless useful for sorting colleges into broad selectivity tiers (Hoxby, 2009).

Results are shown in Table 9. The first two columns re-produce our main results from Table 7. Columns (3) and (4) show results controlling for undergraduate institution selectivity, with the coefficients for Austin less positive in all models but remaining statistically significant in all models where fixed effects are not included (with EOG math only significant at the 10% level).

As a second way to account for selectivity, we add controls for teacher licensure scores in STEM subjects as a measure of individual aptitude.[33] Specifically, we record each teacher's first score on a STEM certification exam and standardize these scores to have mean 0, standard deviation 1 within each subject–year combination and include a cubic function of these scores as an additional control. Results are shown in columns (5) and (6) of Table 9. Relative to our main estimates, coefficients are again attenuated, but even though Austin graduates score substantially above average on certification exams,

[32] We experimented with breaking the selectivity groups into finer groupings (e.g., deciles rather than quintiles) but heaping at certain SAT score cutoffs makes creating these smaller groups of equal sizes difficult.

[33] For math teachers, the standardized (mean 0, standard deviation 1) scores are 0.83 for Austin, 0.53 for replication sites, and 0.07 for non-UTeach teachers. For science, the corresponding scores are 0.81 for Austin, 0.50 for replication sites, and 0.08 for non-UTeach. Overall mean scores are not zero because the standardization process happens before linking to students, so if teachers who score poorly are less likely to enter or persist in teaching, then the analysis sample will have teachers who score above average. This is the same standardization procedure followed by Clotfelter et al. (2010).

estimates of their effectiveness are large even after controlling for certification scores, especially in science.

To probe whether UT Austin graduates' effectiveness can be fully explained by all observable characteristics in our data, we include undergraduate institution selectivity scores, certification scores, and additional controls for certification route. Specifically, we include separate dummy variables for whether a teacher obtained a certification through an alternate route and for whether a teacher is missing in the certification database, as well as an indicator for whether a teacher is STEM certified. Results are shown in Columns (7) and (8), with the UTeach coefficients representing UTeach estimates relative to other teachers who come from standard university-based programs, with similar certification scores, of programs with similar selectivity. With controls for all of these explanatory factors, students of Austin and replication site graduates are estimated to score higher in all subjects, except for replication site graduates in EOC science, by at least 0.05 standard deviations. However, while these effect sizes are still meaningful, they are only statistically significant for replication site graduates in EOC math and Austin graduates in EOC science.[34]

E. Further Exploring Program vs. Institution Effects

In this section, we investigate two additional pieces of evidence in an effort to separate whether the effectiveness of UTeach graduates is due to the UTeach program itself or the fact that UTeach teachers are coming from a select set of TEP institutions. Specifically, for replication sites, we compare the relative performance of UTeach graduates to graduates of the same institutions before UTeach was

[34] Another possible reason the estimated effectiveness of Austin graduates might be higher in science than other sites' is that the Austin program has existed for substantially longer. Thus, it is possible that, once other programs have been in place as long as Austin's, they too will be similarly strong. To investigate whether the other programs show evidence of improving over time, we experimented with program * year interaction terms. Unfortunately, the coefficients for the interaction terms had very large standard errors.

founded. In addition, we estimate the performance of non-UTeach (i.e., non-STEM) graduates of UTeach-affiliated institutions.

We first estimate the effectiveness of STEM teachers who graduated from UTeach replication sites before UTeach was adopted by those sites; unfortunately, we are not able to do the same for Austin graduates prior to UTeach because UTeach has been at Austin for more than 20 years. Specifically, we estimate the coefficients of indicators for whether a teacher graduated from UTeach Austin, from UTeach at a replication site after UTeach was introduced, or from a replication site in the period before UTeach was introduced. Thus, we are comparing the effectiveness of these three groups to average effectiveness of all other teachers in the state.[35] Results are shown in Table 10.

The “Replication Pre” coefficients measure the average effectiveness of STEM teachers who graduated from UTeach institutions prior to the adoption of UTeach relative to teachers from non-UTeach institutions. These estimates range from -0.04 in EOC science to 0.07 in EOC math. Thus, there is some evidence of graduates of replication site institutions being more effective in EOC math, even before UTeach was implemented. However, a comparison of the “Replication Post” to the “Replication Pre” coefficients reveals that replication site UTeach graduates who went through the UTeach program (“Replication Post”) are estimated to be consistently more effective than other STEM teachers who graduated from the same institutions before UTeach was implemented (“Replication Pre”). In all subjects in all models without fixed effects, the pre-UTeach replication site graduates were less effective at raising the test scores of their students than the subsequent UTeach graduates would be by 0.07 to 0.08 standard deviations, although this difference is only statistically significant in some subjects and models. In EOC science, for example, where the replication site coefficient in our main models is not statistically different from zero, the results in Table 10 are suggestive of a large increase in effectiveness

[35] The coefficients for pre-UTeach replication site graduates are identical when not including any UTeach graduates in the regression.

once UTeach was adopted at replication site institutions (from -0.04 to 0.04 standard deviations). This pattern would be consistent with the introduction of UTeach at a campus improving the average quality of STEM teachers who graduate from that university.

As an additional test, we replicate Table 10 on a sample of *non-STEM* (i.e., non-UTeach) teachers from UTeach-affiliated universities. These are teachers who did not earn STEM credentials and are teaching in non-STEM fields. If the improvement in teacher quality at UTeach replication sites associated with the introduction of UTeach in Table 10 were driven by general university-specific trends rather than UTeach itself, one might expect these non-STEM teachers who have graduated from these UTeach-affiliated institutions in the time since UTeach was introduced ("Non-STEM Replication Post") to be larger than the pre-UTeach graduates ("Non-STEM Replication Pre"). Results are shown in Table 11. For EOG reading performance of the replication site graduates, *where non-STEM graduates are not part of the UTeach program at any campus*, we do not see evidence that teachers who graduated from replication sites are more effective in the time period during which UTeach has been operating. Thus, while there is some improvement in EOC reading in the replication sites after the introduction of UTeach in some models, it is not the case that non-STEM teachers at replication sites were uniformly better across all subjects in the more recent UTeach period.

Turning to the Austin coefficient for non-STEM teachers in Table 11, relative to our main results for STEM teachers in Table 7, some of the patterns are broadly similar, with Austin graduates being stronger than other teachers in the state, especially for the more advanced EOC subjects. [36] However, the interpretation of STEM vs. non-STEM results at Austin is complicated by Austin housing the only non-STEM UTeach site: UTeach-Liberal Arts at UT Austin.[37] Thus, Austin is the only campus where a comparison of STEM to non-STEM teachers is not a clean UTeach versus non-UTeach comparison.

---

[36] As with Table 10, we do not estimate pre-UTeach Austin because of the length of time that UTeach has been at Austin.
[37] For more information, see https://liberalarts.utexas.edu/uteach/.

F. Exploring Teacher–Student and Teacher–Class Match Effects

The website for the UTeach site at Austin touts its approach to diversity, in both its training[38] and in producing a more diverse set of teachers than other programs.[39] To investigate whether UTeach graduates are differentially effective at teaching minority students, we include interaction terms between UTeach and student demographics. Results are not shown (for the sake of brevity) but there are few clear patterns with the exception of black students appearing to perform relatively worse when taught by Austin UTeach graduates and FRL-eligible students performing better when taught by replication site graduates.[40] To explore weather UTeach teachers are differentially effective at teaching high-ability classrooms, we compare UTeach teachers in the top 20% of classrooms (as defined by prior math scores) to other teachers in top classrooms. In results available from the authors, we find that students in these high-ability classrooms disproportionately benefit from having either an Austin or UTeach replication site graduate in EOC math subjects, with point estimates on the order of 0.17–0.47, although standard errors are large in some cases.

VI. Conclusion

Relative to other teachers in the state, we find that graduates of both the UTeach founding site at Austin and the replication sites are more effective as measured by their ability to raise student test scores in math. In science, we find that Austin graduates improve the test scores of their students, while

[38] https://institute.uteach.utexas.edu/sessions/three-university-perspectives-weaving-equity-diversity-and-current-issues-classroom

[39]http://www.sheeo.org/sites/default/files/Weds%201115%20Marcus%20Lingenfelter%202013UTeachGradReportFinal120204.pdf

[40] When interacting UTeach with school characteristics, we find suggestive evidence that increases in the school share of black students is associated with reductions in the effectiveness of Austin grads, but these estimates are very imprecise. These results are available from the authors upon request.

estimates for replication sites are smaller and not statistically significant. In some cases, these effect sizes are very large, with high school math and science students taught by UTeach Austin graduates estimated to accrue 4-6 months of additional learning in a 9 month school year. We conduct several tests to assess whether the positive UTeach effects we find might be driven by overall university-level effects such as the selection of students to universities rather than UTeach program effects. While it is difficult to definitively distinguish between true program effects and general institution-level effects, these tests suggest that the introduction of UTeach to a given university is associated with an increase in the performance of the STEM teachers produced by that university. In addition, the inclusion of proxies for selectivity such as the SAT scores of incoming students cannot fully explain the UTeach effects we find.

While our tests are suggestive of true UTeach program effects, there are several reasons that the findings in this paper are still important even if fully driven by institution effects. First, a primary goal of UTeach is to increase the production of STEM teachers from partner universities. While not the focus of the paper, we find evidence that UTeach partner universities do produce more STEM teachers after the implementation of UTeach. Thus, even if the UTeach program itself does not improve the quality of a given teacher candidate, by producing more teachers from universities with above average teachers, the program would still improve the overall quality of the STEM workforce. Second, our results suggest that whether a teacher candidate graduated from a UTeach program is a signal of an applicant's quality. While distinguishing university selection effects from training effects is important for policymakers who operate at the state and university levels, for leaders of secondary schools interested in the learning of their students, the question of whether graduates of a given university are more effective because of the selection process into the university or because of the training received at that university is not particularly relevant. Finally, UTeach offers a 4 year degree plan that condenses the certification courses that were offered in the previous programs at UTeach-affiliated institutions prior to UTeach implementation. Our results suggest that condensing these courses has not resulted in detrimental performance once teachers enter the classroom.

References

Ares, N., Stroup, W. M., & Schademan, A. R. (2009). The power of mediating artifacts in group-level development of mathematical discourses. *Cognition and Instruction*, *27*(1), 1–24.

Augustine, N. R. (2007). Is America falling off the flat earth?. National Academies Press.

Bacher-Hicks, A., Kane, T. J., & Staiger, D. O. (2014). *Validating teacher effect estimates using changes in teacher assignments in Los Angeles* (No. w20657). Cambridge, MA: National Bureau of Economic Research.

Borman, G. D., & Dowling, N. M. (2008). Teacher attrition and retention: A meta-analytic and narrative review of the research. *Review of Educational Research*, *78*(3), 367–409.

Boyd, D., Grossman, P., Lankford, H., Loeb, S., & Wyckoff, J. (2006). How changes in entry requirements alter the teacher workforce and affect student achievement. *Education Finance and Policy, 1*(2), 176–216.

Boyd, D., Lankford, H., Loeb, P. & Wyckoff, J. (2005). The draw of home: How teachers' preferences for proximity disadvantage urban youth. *Journal of Policy Analysis and Management*, *24*(1), 113–132.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014a). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *The American Economic Review*, *104*(9), 2593–2632.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014b). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *The American Economic Review*, *104*(9), 2633–2679.

Clark, M. A., Chiang, H. S., Silva, T., McConnell, S., Sonnenfeld, K., Erbe, A., & Puma, M. (2013). The effectiveness of secondary math teachers from Teach For America and the Teaching Fellows programs. Princeton, NJ: Mathematica Policy Research.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review*, *26*(6), 673–682.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2010). Teacher credentials and student achievement in high school a cross-subject analysis with student fixed effects. *Journal of Human Resources*, *45*(3), 655–681.

Confrey, J., Makar, K., & Kazak, S. (2004). Undertaking data analysis of student outcomes as professional development for teachers. *ZDM: International Reviews on Mathematics Education*, *36*(1), 32–40.

Cowan, J., Goldhaber, D., Hayes, K., & Theobald, R. (2016). Missing elements in the discussion of teacher shortages. *Educational Researcher*, 0013189X16679145.

Dee, T. S. (2004). Teachers, race, and student achievement in a randomized experiment. *Review of Economics and Statistics*, *86*(1), 195–210.

Dickinson, G., & Summers, E. J. (2010). (Re)Anchored, video-centered engagement: The transferability of preservice training to practice. *Contemporary Issues in Technology and Teacher Education*, *10*(1), 106–118.

Goldhaber, D. (2008). Teachers matter, but effective teacher quality policies are elusive. In *Handbook of research in education finance and policy*. (pp. 146–165).

Goldhaber, D. D., & Brewer, D. J. (1997). Why don't schools and teachers seem to matter? Assessing the impact of unobservables on educational productivity. *Journal of Human Resources*, 505-523.

Goldhaber, D., Gratz, T., & Theobald, R. (2016). What's in a Teacher Test? Assessing the Relationship Between Teacher Licensure Test Scores and Student STEM Achievement and Course-Taking. *Working paper.*

Goldhaber, D., Theobald, R., & Fumia, D. (2018). Teacher Quality Gaps and Student Outcomes: Assessing the Association Between Teacher Assignments and Student Math Test Scores and High School Course Taking. *CALDER Working Paper #185*.

Goldhaber, D., & Hansen, M. (2010). Race, gender, and teacher testing: How informative a tool is teacher licensure testing? *American Educational Research Journal*, *47*(1), 218–251.

Goldhaber, D., Liddle, S., & Theobald, R. (2013). The gateway to the profession: Assessing teacher preparation programs based on student achievement. *Economics of Education Review*, *34*, 29–44.

Goldhaber, D., & Walch, J. (2013). *Rhetoric versus reality: Is the academic caliber of the teacher workforce changing?* (Vol. 4). CEDR Working Paper 2013.Hansen, M., Backes, B., Brady, V., & Xu, Z. (2015). *Examining spillover effects from Teach For America corps members in Miami-Dade County Public Schools*. National Center for Analysis of Longitudinal Data in Education Research Working Paper, 113.

Hanushek, E. A., Kain, J. F., O'Brien, D. M., & Rivkin, S. G. (2005). *The market for teacher quality* (No. w11154). Cambridge, MA: National Bureau of Economic Research.

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *The American Economic Review*, *100*(2), 267–271.

Harris, D., & Sass, T. (2006). *The effects of teacher training on teacher value-added*. Florida State University.

Heck, R. H. (2009). Teacher effectiveness and student achievement: Investigating a multilevel cross-classified model. *Journal of Educational Administration*, *47*(2), 227–249.

Hock, H., & Isenberg, E. (2012). Methods for Accounting for Co-Teaching in Value-Added Models. *Working Paper*. Mathematica Policy Research, Inc.

Holdren, J. P., Lander, E. S., & Varmus, H. (2010). Prepare and inspire: K–12 education in science, technology, engineering, and math (STEM) for America’s future. *Executive Report).* Washington, DC: President's Council of Advisors on Science and Technology.

Hoxby, C. M. (2009). The changing selectivity of American colleges. *The Journal of Economic Perspectives*, *23*(4), 95–118.

Ingersoll, R. M. (2001). Teacher turnover and teacher shortages: An organizational analysis. *American educational research journal*, *38*(3), 499–534.

Jackson, C. K. (2014). Teacher quality at the high-school level: The importance of accounting for tracks. *Journal of Labor Economics*, 32(4), 645-684.

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review, 27*(6), 615–631.

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* (No. w14607). Cambridge, MA: National Bureau of Economic Research.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). Have we identified effective teachers? Validating measures of effective teaching using random assignment. In *Research Paper. MET Project.* Seattle, WA: Bill & Melinda Gates Foundation.

Koedel, C., Parsons, E., Podgursky, M., & Ehlert, M. (2015). Teacher preparation programs and teacher quality: Are there real differences across programs? *Education Finance and Policy*. http://www.mitpressjournals.org/doi/pdf/10.1162/EDFP_a_00172.

Lipsey, Mark W., Kelly Puzio, Cathy Yun, Michael A. Hebert, Kasia Steinka-Fry, Mikel W. Cole, Megan Roberts, Karen S. Anthony, and Matthew D. Busick. (2012). "Translating the Statistical Representation of the Effects of Education Interventions into More Readily Interpretable Forms." National Center for Special Education Research.

Makar, K., & Confrey, J. (2005). "Variation-talk": Articulating meaning in statistics. *Statistics Education Research Journal*, *4*(1), 27–54.

Marder, M., & Walkington, C. (2014). Classroom observation and value-added models give complementary information about the quality of mathematics teaching. In T. Kane, K. Kerr, & R. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the Measuring Effective Teaching project* (pp. 234–277). New York, NY: Wiley.

Marshall, J., & Young, E. S. (2006). Preservice teachers' theory development in physical and simulated environments. *Journal of Research in Science Teaching*, *43*(9), 907–937.

Peterson, P. E., Woessmann, L., Hanushek, E. A., & Lastra-Anadón, C. X. (2011). *Globally challenged: Are U.S. students ready to compete? The latest on each state's international standing in math and reading*. PEPG 11-03. Program on Education Policy and Governance, Harvard University.

Rice, J. K., & Schwartz, A. E. (2008). Toward an understanding of productivity in education. In *Handbook of research in education finance and policy* (pp. 131–145).

Rockoff, J. E., Jacob, B. A., Kane, T. J., & Staiger, D. O. (2011). Can you recognize an effective teacher when you recruit one? *Education*, *6*(1), 43–74.

Stroup, W., Hills, T., & Carmona, G. (2011). Computing the average square: An agent-based introduction to aspects of psychometric practice. *Technology, Knowledge and Learning*, *16*(3), 199–220.

Von Hippel, P. T., Bellows, L., Osborne, C., Lincove, J. A., & Mills, N. (2016). Teacher quality differences between teacher preparation programs: How big? How reliable? Which programs are different? *Economics of Education Review*, *53*, 31–45.

West, M. (2013). Do Math and Science Teachers Earn More Outside of Education? *Brookings*. https://www.brookings.edu/research/do-math-and-science-teachers-earn-more-outside-of-education/

Xu, Z., Hannaway, J., & Taylor, C. (2011). Making a difference? The effects of Teach For America in high school. *Journal of Policy Analysis and Management*, *30*(3), 447–469.

Xu, Z., Ozek, U., & Corritore, M. (2012). Portability of Teacher Effectiveness across School Settings. Working Paper 77. *National Center for Analysis of Longitudinal Data in Education Research*.

Tables

Table 1: Number of UTeach Teachers by Campus and Sample Year

	EOG M					EOC M					EOC S				
	2012	2013	2014	2015	2016	2012	2013	2014	2015	2016	2012	2013	2014	2015	2016
Arlington	0	0	0	0	≤5	0	0	0	9	26	0	0	0	8	14
Austin	17	16	14	17	18	115	131	113	114	120	80	97	75	88	82
Dallas	0	0	0	≤5	≤5	≤5	≤5	12	17	20	0	≤5	10	14	21
Houston	≤5	≤5	≤5	7	6	12	24	28	43	51	≤5	14	23	32	43
UNT	≤5	≤5	≤5	≤5	11	12	38	56	68	80	≤5	6	8	18	25
Rio Grande	0	0	0	11	22	0	0	0	10	23	0	0	0	≤5	16
Tyler	0	0	≤5	18	28	0	0	≤5	8	13	0	0	0	0	≤5

Notes: Cell values indicate the number of unique teachers linked to students with valid test scores. "EOG M" denotes teachers of EOG math (Grades 6–8), "EOC M" teachers of EOC math (Algebra I, Geometry, and Algebra II), and "EOC S" denotes teachers of EOC science (Biology, Chemistry, and Physics).

Table 2: Number of Teachers in Analysis Sample by Campus and Graduation Year

	Arlington	Austin	Dallas	Houston	Rio Grande	Tyler	UNT
1995	17	9	13	17	63	13	20
1996	25	12	14	15	62	7	15
1997	13	10	6	11	53	7	20
1998	12	7	13	14	55	≤5	14
1999	8	8	13	12	60	5	22
2000	14	9	5	11	40	8	20
2001	6	18	11	14	49	11	23
2002	5	14	7	7	54	13	15
2003	6	15	≤5	8	43	6	7
2004	6	16	≤5	≤5	42	11	12
2005	9	33	7	12	45	5	18
2006	8	25	9	13	39	10	18
2007	5	26	8	10	49	7	19
2008	6	43	≤5	5	38	7	19
2009	10	40	5	5	39	10	13
2010	11	27	≤5	10	47	10	30
2011	19	47	8	11	39	10	29
2012	≤5	43	11	27	32	23	38
2013	8	37	13	27	25	21	33
2014	26	40	16	36	42	18	30
2015	27	28	9	22	24	13	24

Notes: Each cell denotes the number of STEM teachers in our analysis sample who graduated from a given campus in a given calendar year. Numbers below horizontal lines represent teachers who graduated after UTeach had been implemented at a given campus.

Table 3: Summary Statistics of Students Taught by UTeach and Non-UTeach Teachers

	EOG M			EOC M			EOC S		
	Non-U	Austin	Replication	Non-U	Austin	Replication	Non-U	Austin	Replication
Male	0.51	0.47	0.50	0.50	0.51	0.50	0.49	0.50	0.50
Black	0.11	0.10	0.10	0.12	0.12	0.16	0.12	0.11	0.16
Hispanic	0.52	0.36	0.61	0.48	0.50	0.52	0.48	0.42	0.55
Asian	0.03	0.09	0.04	0.04	0.06	0.05	0.04	0.06	0.05
White	0.31	0.42	0.23	0.32	0.29	0.23	0.32	0.38	0.20
Other	0.02	0.03	0.00	0.02	0.02	0.02	0.02	0.03	0.02
LEP	0.16	0.08	0.24	0.09	0.12	0.16	0.09	0.09	0.17
FRL	0.58	0.38	0.67	0.51	0.50	0.58	0.51	0.41	0.58
Spec Ed	0.05	0.03	0.04	0.05	0.04	0.05	0.04	0.04	0.04
Gifted	0.09	0.15	0.11	0.12	0.11	0.07	0.11	0.13	0.08
Grade	6.93	7.48	7.20	8.99	9.13	9.08	9.27	9.33	9.16
	(0.77)	(0.69)	(0.74)	(0.65)	(0.58)	(0.47)	(0.51)	(0.52)	(0.35)
Lag m	-0.04	0.36	-0.17	0.15	0.14	-0.05	-0.01	0.17	-0.13
	(0.51)	(0.52)	(0.48)	(0.72)	(0.73)	(0.65)	(0.51)	(0.54)	(0.45)
Lag r	-0.02	0.43	-0.11	0.08	0.07	-0.14	0.05	0.24	-0.15
	(0.46)	(0.50)	(0.48)	(0.60)	(0.62)	(0.53)	(0.54)	(0.57)	(0.51)
Lag s							0.02	0.28	-0.10
							(0.60)	(0.65)	(0.56)
Miss lagm	0.04	0.04	0.04	0.05	0.05	0.05	0.06	0.06	0.06
Miss lagr	0.04	0.04	0.03	0.06	0.06	0.06	0.07	0.07	0.06
Miss prior s							0.07	0.07	0.08

Notes: “Non-U” denotes non-UTeach, “Austin” denotes the UT Austin site, and “Replication” denotes UTeach replication (non-Austin) sites. “EOG M” denotes teachers of EOG math (Grades 6–8), “EOC M” teachers of EOC math (Algebra I, Geometry, and Algebra II), and “EOC S” denotes teachers of EOC science (Biology, Chemistry, and Physics). Standard deviations in parentheses.

Table 4: Summary Statistics for Teachers

	EOG M			EOC M			EOC S		
	Non-U	Austin	Replication	Non-U	Austin	Replication	Non-U	Austin	Replication
Yrs exp	9.45	5.05	0.71	10.67	4.01	1.05	10.28	3.98	0.76
	(8.62)	(3.64)	(0.85)	(9.37)	(3.45)	(1.05)	(9.39)	(3.16)	(0.95)
1st yr teacher	0.09	0.09	0.41	0.08	0.15	0.33	0.09	0.11	0.44
2-3rd yr teacher	0.14	0.20	0.41	0.13	0.25	0.46	0.13	0.26	0.38
Missing tch exp	0.03	0.04	0.15	0.03	0.06	0.11	0.03	0.07	0.12
SAT 75th pct M	584	720	591	588	720	622	590	720	634
	(61)		(36)	(64)		(35)	(65)		(41)
SAT 75th pct R	568	690	567	571	690	601	573	690	608
	(57)		(35)	(59)		(32)	(59)		(39)
Missing SAT M scores	0.25	0	0	0.29	0	0	0.27	0	0
STEM certified	0.63	1	1	0.91	1	1	0.95	1	1
Black	0.12	0.05	0.07	0.09	0.09	0.06	0.09	0.01	0.06
Hispanic	0.20	0.34	0.35	0.20	0.33	0.31	0.19	0.16	0.27
White	0.63	0.52	0.46	0.65	0.49	0.49	0.65	0.60	0.45
Teacher-year obs	69318	82	136	60208	593	557	46450	422	269

Notes: “Non-U” denotes non-UTeach, “Austin” denotes the UT Austin site, and “Replication” denotes UTeach replication (non-Austin) sites. “EOG M” denotes teachers of EOG math (Grades 6–8), “EOC M” teachers of EOC math (Algebra I, Geometry, and Algebra II), and “EOC S” denotes teachers of EOC science (Biology, Chemistry, and Physics). Standard deviations in parentheses.

Table 5: Coefficient Estimates of UTeach by Subject, Math

	1	2	3	4
Panel 1: EOG Math (unique UTeach: 134)				
UTeach	0.06	0.10**	0.08*	0.08**
	(0.04)	(0.04)	(0.04)	(0.04)
Panel 2: Algebra I (unique UTeach: 474)				
UTeach	0.12***	0.13***	0.12***	0.01
	(0.03)	(0.03)	(0.03)	(0.01)
Panel 3: Geometry (unique UTeach: 164)				
UTeach	0.06*	0.08**	0.07**	0.04
	(0.04)	(0.04)	(0.04)	(0.03)
Panel 4: Algebra II (unique UTeach: 113)				
UTeach	0.05	0.09*	0.09*	0.14***
	(0.05)	(0.05)	(0.05)	(0.05)
Student chars	X	X	X	X
Teacher chars		X	X	X
STEM certified			X	
School fixed effect				X

Notes: Additional controls include teacher experience, a dummy variable for missing teacher experience, identifiers for FRL, gifted, LEP, and special education, cubic functions of prior reading and prior match scores, and indicators for missing test scores. Standard errors displayed in parentheses are clustered at the school-cohort level. Coefficients represent effect sizes in standard deviations.

Table 6: Coefficient Estimates of UTeach by Subject, Science

	1	2	3	4
Panel 1: Biology (unique UTeach: 296)				
UTeach	0.07***	0.09***	0.09***	0.01
	(0.02)	(0.02)	(0.02)	(0.02)
Panel 2: Chemistry (unique UTeach: 104)				
UTeach	0.11**	0.12***	0.12***	0.05*
	(0.05)	(0.04)	(0.04)	(0.03)
Panel 3: Physics (unique UTeach: 51)				
UTeach	0.10	0.15	0.15	0.08
	(0.09)	(0.09)	(0.09)	(0.08)
Student chars	X	X	X	X
Teacher chars		X	X	X
STEM certified			X	
School fixed effect				X

Notes: See notes from Table 5.

Table 7: Coefficient Estimates for UT Austin and UTeach Replication Sites

	1	2	3	4
Panel 1: EOG Math				
Austin	0.12**	0.13**	0.12**	0.07*
	(0.05)	(0.05)	(0.05)	(0.04)
Replication	0.02	0.08	0.06	0.08
	(0.06)	(0.06)	(0.06)	(0.05)
Panel 2: EOC Math				
Austin	0.10***	0.11***	0.10***	-0.00
	(0.04)	(0.04)	(0.04)	(0.02)
Replication	0.12***	0.14***	0.13***	0.01
	(0.03)	(0.03)	(0.03)	(0.02)
Panel 3: EOC Science				
Austin	0.13***	0.13***	0.13***	0.04**
	(0.03)	(0.02)	(0.02)	(0.02)
Replication	0.01	0.04	0.04	-0.03
	(0.03)	(0.03)	(0.03)	(0.03)
Student chars	X	X	X	X
Teacher chars		X	X	X
STEM certified			X	
School fixed effect				X

Notes: See notes from Table 5. “EOG M” denotes teachers of EOG math (Grades 6–8), “EOC M” teachers of EOC math (Algebra I, Geometry, and Algebra II), and “EOC S” denotes teachers of EOC science (Biology, Chemistry, and Physics).

Table 8: Specification and robustness checks

	EOG M		EOC M		EOC S	
	Austin	Replication	Austin	Replication	Austin	Replication
	1	2	3	4	5	6
1: Main OLS results (Table 7)	0.13**	0.08	0.11***	0.14***	0.13***	0.04
	(0.05)	(0.06)	(0.04)	(0.03)	(0.02)	(0.03)
2: Main school FE results (Table 7)	0.07*	0.08	-0.00	0.01	0.04**	-0.03
	(0.04)	(0.05)	(0.02)	(0.02)	(0.02)	(0.03)
3: School-track FE	0.04	0.01	0.01	0.01	0.04***	-0.02
	(0.03)	(0.06)	(0.02)	(0.02)	(0.02)	(0.01)
4: School characteristics controls	0.08	0.06	0.09***	0.12***	0.09***	0.02
	(0.05)	(0.05)	(0.04)	(0.03)	(0.03)	(0.03)
5: School + class chars controls	0.06	0.05	0.09***	0.10***	0.08***	0.03
	(0.05)	(0.04)	(0.03)	(0.03)	(0.02)	(0.03)
6: Schools w/o strong tracking (OLS)	0.07	0.03	0.06	0.16***	0.14***	0.04
	(0.05)	(0.03)	(0.04)	(0.04)	(0.03)	(0.03)
7: Schools w/o strong tracking (FE)	0.01	0.02	-0.02	0.01	0.06***	-0.02
	(0.03)	(0.05)	(0.02)	(0.02)	(0.02)	(0.03)
8: Novice teachers only (yrs 1-3) (OLS)	0.08	0.06	0.13***	0.15***	0.16***	0.04
	(0.07)	(0.04)	(0.04)	(0.03)	(0.04)	(0.03)
9: Novice teachers only (yrs 1-3) (FE)	0.10*	0.07	-0.01	0.02	0.07**	-0.00
	(0.05)	(0.04)	(0.03)	(0.02)	(0.03)	(0.03)
10: Excluding 1st UTeach cohort (OLS)	0.09*	0.06	0.11***	0.15***	0.13***	0.03
	(0.05)	(0.03)	(0.04)	(0.04)	(0.02)	(0.03)
11: Excluding 1st UTeach cohort (FE)	0.04	0.04*	0	0.01	0.04**	-0.04
	(0.04)	(0.03)	(0.02)	(0.02)	(0.02)	(0.03)

Notes: See notes from Table 5. “EOG M” denotes teachers of EOG math (Grades 6–8), “EOC M” teachers of EOC math (Algebra I, Geometry, and Algebra II), and “EOC S” denotes teachers of EOC science (Biology, Chemistry, and Physics). The controls used for school and class characteristics in rows 4 and 5 include cubic polynomials in percent black, percent Hispanic, and prior math and reading scores. In rows 6 and 7, we remove the approximately 20% of schools where the average deviation in prior math scores at the classroom level from the overall school average exceeds 0.05 standard deviations.

Table 9: Coefficient Estimates for UTeach When Controlling for Measures of Selection

	1	2	3	4	5	6	7	8
Panel 1: EOG Math								
Austin	0.13**	0.07*	0.09*	0.06	0.09*	0.04	0.06	0.04
	(0.05)	(0.04)	(0.05)	(0.04)	(0.05)	(0.04)	(0.05)	(0.04)
Replication	0.08	0.08	0.10	0.08	0.05	0.05	0.07	0.06
	(0.06)	(0.05)	(0.06)	(0.05)	(0.06)	(0.05)	(0.06)	(0.05)
Panel 2: EOC Math								
Austin	0.11***	-0.00	0.08**	-0.01	0.07*	-0.02	0.05	-0.03
	(0.04)	(0.02)	(0.04)	(0.02)	(0.04)	(0.02)	(0.04)	(0.02)
Replication	0.14***	0.01	0.15***	0.02	0.10***	-0.00	0.11***	0.00
	(0.03)	(0.02)	(0.03)	(0.02)	(0.03)	(0.02)	(0.03)	(0.02)
Panel 3: EOC Science								
Austin	0.13***	0.04**	0.07***	0.02	0.11***	0.04*	0.06**	0.02
	(0.02)	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)	(0.03)	(0.02)
Replication	0.04	-0.03	0.01	-0.04*	0.03	-0.03	-0.00	-0.04*
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.02)	(0.03)	(0.03)
Measures of undergrad selectivity			X	X			X	X
Teacher certification score					X	X	X	X
Certification type							X	X
School fixed effect		X		X		X		X

Notes: See notes from Table 5. "EOG Math" denotes teachers of EOG math (Grades 6–8), "EOC Math" teachers of EOC math (Algebra I, Geometry, and Algebra II), and "EOC Science" denotes teachers of EOC science (Biology, Chemistry, and Physics). Undergraduate selectivity controls include the 75th percentile of the SAT math and reading scores of incoming students, certification score controls include cubic functions of a teacher's first STEM certification score, and certification type controls include an indicator for STEM certified as well as whether the teacher came from an alternate certification route or is missing from the certification database.

Table 10: Coefficient Estimates for UTeach and Replication Campuses Prior to UTeach

	EOG M		EOC M		EOC S	
	1	2	3	4	5	6
Austin	0.13**	0.07*	0.11***	-0.00	0.13***	0.04
	(0.05)	(0.04)	(0.04)	(0.02)	(0.02)	(0.02)
Replication Pre	0.01	0.03***	0.07***	0.02	-0.04***	-0.01
	(0.01)	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)
Replication Post	0.08	0.08	0.14***	0.01	0.04	-0.03
	(0.05)	(0.06)	(0.03)	(0.02)	(0.03)	(0.03)
School FE		X		X		X

Notes: See notes from Table 5. “EOG M” denotes teachers of EOG math (Grades 6–8), “EOC M” teachers of EOC math (Algebra I, Geometry, and Algebra II), and “EOC S” denotes teachers of EOC science (Biology, Chemistry, and Physics). “Replication Pre” denotes teachers who were trained at UTeach replication site institutions prior to the implementation of UTeach and thus are not UTeach teachers, and “Replication Post” denotes teachers who were trained at UTeach institutions after the introduction of UTeach.

Table 11: Reading Results for non-STEM Graduates of Campuses With UTeach

	1	2	3
Panel 1: EOG Reading			
Non-STEM Austin	0.03***	0.04***	0.01
	(0.01)	(0.01)	(0.01)
Non-STEM Replication Post	-0.02**	-0.01	-0.02**
	(0.01)	(0.01)	(0.01)
Non-STEM Replication Pre	-0.01*	-0.01**	-0.00
	(0.00)	(0.00)	(0.00)
Panel 2: EOC Reading			
Non-STEM Austin	0.09***	0.11***	0.02*
	(0.02)	(0.02)	(0.01)
Non-STEM Replication Post	0.04**	0.07***	0.02**
	(0.01)	(0.01)	(0.01)
Non-STEM Replication Pre	0.03***	0.03***	0.02***
	(0.01)	(0.01)	(0.01)
Student chars	X	X	X
Teacher chars		X	X
Fixed effect			School

Notes: See notes from Table 5. In any given year, only one EOC at a given level is administered (e.g., English I or Reading I). "EOG R" denotes teachers of EOG reading (Grades 6_8) and "EOC R" teachers of EOC reading (Reading II, Reading II, English I, and English II). "Non-STEM Austin" denotes non-STEM teachers who were trained at UT Austin, "Non-STEM Replication Pre" denotes teachers who were trained at UTeach replication site institutions prior to the implementation of UTeach, and "Non-STEM Replication Post" denotes non-STEM teachers who were trained at UTeach institutions after the introduction of UTeach. The teachers in this table do not have STEM certificates and do not teach in STEM fields.

Appendix

Appendix Table 1: Full Regression Coefficients From Most Commonly Taken Subjects

	EOG M	Algebra I	Biology
	(1)	(2)	(3)
UTeach	0.10**	0.13***	0.09***
	(0.04)	(0.03)	(0.02)
Male	0.02***	-0.01***	0.02***
	(0.00)	(0.00)	(0.00)
Black	-0.12***	-0.09***	-0.07***
	(0.00)	(0.01)	(0.00)
Hispanic	-0.07***	-0.04***	-0.08***
	(0.00)	(0.01)	(0.00)
Asian	0.36***	0.16***	0.25***
	(0.01)	(0.01)	(0.01)
Other	0	0.01	0.03***
	(0.00)	(0.01)	0.00
Prior Math	0.69***	0.43***	0.14***
	(0.00)	(0.00)	(0.00)
Prior reading	0.17***	0.14***	0.29***
	(0.00)	(0.00)	(0.00)
Prior science			0.44***
			(0.00)
1 year exp	0.04***	0.07***	0.04***
	0.00	(0.01)	(0.01)
2 years exp	0.07***	0.07***	0.06***
	(0.01)	(0.01)	(0.01)
3 years exp	0.08***	0.09***	0.08***
	(0.01)	(0.01)	(0.01)
4 years exp	0.09***	0.09***	0.08***
	(0.01)	(0.01)	(0.01)
5 years exp	0.09***	0.07***	0.07***
	(0.01)	(0.01)	(0.01)
6 years exp	0.09***	0.05***	0.07***
	(0.01)	(0.02)	(0.01)
7 years exp	0.10***	0.05***	0.07***
	(0.01)	(0.02)	(0.01)
8 years exp	0.10***	0.08***	0.07***
	(0.01)	(0.02)	(0.01)
9 years exp	0.10***	0.06***	0.06***
	(0.01)	(0.02)	(0.01)
10 years exp	0.11***	0.08***	0.08***

	(0.01)	(0.02)	(0.01)
Over 10 years exp	0.09***	0.06***	0.05***
	0.00	(0.01)	(0.01)
Missing exp	-0.01	-0.11***	-0.04***
	(0.01)	(0.03)	(0.01)
Student chars	X	X	X
Teacher chars	X	X	X

Notes: Additional controls include squared and cubic terms in prior test scores, limited English proficiency, FRL eligibility, special education status, and gifted status. Standard errors clustered at the school-cohort level. The regressions represented here are identical to those that generated column 2 of Tables 5 and 6.

Appendix Table 2: Measures of Sorting to UTeach Schools, non-UTeach Teachers

	EOG M 1	Algebra I 2	Geometry 3	Algebra II 4	Biology 5	Chemistry 6	Physics 7
Panel 1: Certification score							
School UTeach	0.05*** (0.02)	0.11*** (0.02)	0.16*** (0.04)	0.07** (0.03)	0.10*** (0.02)	0.14*** (0.04)	0.13* (0.07)
Panel 2: Teacher ever taught in UTeach school (sample = non-UTeach schools)							
Teacher ever in UTeach school	0.02 (0.01)	0.08*** (0.02)			-0.01 (0.01)		

Notes: Panel 1 displays the results of a student-level regression where the outcome variable is a student's teacher's certification score, conditional on student demographic characteristics and prior achievement. Each coefficient is the estimated change associated with a student's presence in a school that ever has a UTeach teacher in a given subject. Panel 2 displays a regression of student test scores on demographic characteristics, prior test scores, and an indicator for whether the student is being taught by a teacher who is ever observed teaching in a UTeach school, with the regression performed on a sample of students and teachers in the years they are not in UTeach schools.

Appendix Table 3: Program Estimates

	EOG M		EOC M		EOC S	
	1	2	3	4	5	6
Arlington	0.17	0.08	0.06	0.06	0.07	0.03
	(0.15)	(0.13)	(0.07)	(0.07)	(0.06)	(0.04)
Austin	0.13**	0.07*	0.11***	-0.00	0.13***	0.04**
	(0.05)	(0.04)	(0.04)	(0.02)	(0.02)	(0.02)
Tyler	-0.03	-0.02	-0.06	-0.09	0.13	0.25***
	(0.05)	(0.04)	(0.15)	(0.10)	(0.13)	(0.03)
Dallas	-0.05***	-0.12	0.21***	0.01	0.21*	0.08
	(0.01)	(0.08)	(0.05)	(0.04)	(0.12)	(0.11)
RioGrande	-0.01	0.01	0.14	0.11	-0.07*	-0.03
	(0.07)	(0.06)	(0.12)	(0.08)	(0.04)	(0.04)
UNT	0.03	0.07**	0.11***	-0.01	0.01	-0.00
	(0.04)	(0.03)	(0.04)	(0.03)	(0.03)	(0.03)
Houston	0.45*	0.36	0.21***	0.04	0.00	-0.09***
	(0.25)	(0.22)	(0.06)	(0.04)	(0.03)	(0.03)
School FEs		x		x		X

Notes: “EOG M” denotes teachers of EOG math (Grades 6–8), “EOC M” teachers of EOC math (Algebra I, Geometry, and Algebra II), and “EOC S” denotes teachers of EOC science (Biology, Chemistry, and Physics).